News Narrative Modeling via Event Graphs

Celine Lim Shi-Yen
A0171682B
e0201252@u.nus.edu

Clare Yeo Kai Wen
A0157555W
e0036068@u.nus.edu

Le Tan Dang Khoa
A0274718U
e1124464@u.nus.edu

Thet Su Win
A0304980A
e1374068@u.nus.edu

***Abstract*—We present a modular framework for extracting and linking events from unstructured news to model narrative progression. Our approach combines large language models for zero-shot event extraction, unsupervised clustering for grouping semantically similar mentions, and weakly supervised methods for inferring temporal and causal links. The resulting event-centric graph captures inter-event dependencies and supports applications such as story tracking, policy impact analysis, and knowledge graph construction. Graph-based metrics like PageRank and betweenness centrality highlight key trigger points. Qualitative evaluation shows the system organizes open-domain news into interpretable structures.**

I. Introduction

Understanding how events occur, change, and affect one another is important for analyzing complex global issues. Areas such as international politics, economic policy, and national security often involve chains of related events that appear across many sources and time periods. Extracting and organizing these events in a structured format helps support timely analysis, historical review, and forecasting. This is particularly relevant in geopolitical contexts, where actions such as trade decisions or diplomatic negotiations often lead to cascading effects across sectors.

News articles provide frequent updates about such events, but their unstructured format makes information extraction difficult. The same event may be described in different ways, spread across several articles, and appear in broader narratives that make their relationships unclear. These challenges make it hard to build consistent event representations and limit the use of traditional extraction methods, which often rely on fixed schemas or annotated training data.

To address these challenges, we propose a modular framework for extracting, clustering, and linking events from open-domain news. The system has three main components. First, a large language model (LLM) is used to extract structured events in a zero-shot manner. Second, semantically similar event mentions are grouped using unsupervised clustering. Third, we infer temporal and causal relations between event clusters using a combination of commonsense reasoning and semantic similarity. The resulting structure is an event-centric knowledge graph, where nodes represent real-world events and edges encode inter-event dependencies.

The proposed framework can support many real-world applications. Journalists and analysts can use it to follow how stories develop across sources and time. Policymakers and economists can study the downstream impact of specific decisions or identify patterns that precede large-scale disruptions. Intelligence analysts can track event cascades to identify signs of rising geopolitical risk. Financial institutions and businesses can explore historical event chains that align with market movements or disruptions in global supply networks.

By structuring these events as a knowledge graph, the system enables query-based exploration and structural reasoning. Users can pose questions such as: *"What events usually precede the collapse of trade negotiations?"*, or *"Which developments have led to inflationary trends in the past?"* Centrality metrics like PageRank and betweenness help surface key trigger points or bridging events, revealing how specific developments influence or connect broader narratives. An interactive graph interface further supports visual and analytical engagement with these patterns.

While our experiments are scoped to the US-China trade narrative for demonstration purposes, the design of the framework itself is domain-agnostic and scalable by construction. The use of zero-shot event extraction with large language models, unsupervised clustering, and weakly supervised causal inference are all general techniques that are not restricted to any specific topic. Our choice to focus on US-China trade reflects a strategic scoping decision rather than a limitation, allowing us to provide a grounded, interpretable case study of how the framework can handle real-world, multi-domain geopolitical complexity.

The framework is adaptable to other domains such as climate policy, regional conflicts, or financial regulation with minimal changes. Applying the method to a new topic only requires replacing the input corpus with domain-relevant articles. Although the current dataset is focused on a specific theme, the overall approach remains modular, generalizable, and scalable across different subject areas and use cases.

II. Dataset

To support the development of our event-based framework for geopolitical analysis, we construct a targeted news dataset focused on economic and international policy topics. The data is collected from two major Southeast Asian media outlets: The Straits Times (ST)[1] and Channel News Asia (CNA)[2]. These sources are known for their consistent reporting on regional and global developments, particularly in areas related to trade, diplomacy, and political decision-making. Their detailed coverage makes them a strong fit for tasks involving event extraction and causal inference.

[1] https://www.straitstimes.com
[2] https://www.channelnewsasia.com

Articles were collected over a 12-month period from February 2024 to February 2025. To focus our analysis on a coherent geopolitical context, we applied a keyword filter requiring the presence of "trade", "US", and "China" in the article text. This filter was not a technical constraint, but a deliberate scoping decision to center the project on a real-world narrative of high strategic relevance: US-China trade relations. This framing enabled us to construct a more interpretable event graph while still capturing diverse event types, such as policy announcements, economic retaliation, military signaling, and diplomatic meetings.

The articles exhibit a formal and factual writing style that highlights specific actors, actions, and temporal references, making them well-suited for structured event extraction. While the dataset includes a variety of article lengths and formats, this variability reflects the natural complexity of real-world reporting and enhances the generalizability of our methods beyond stylized or synthetic datasets. The design of our system remains modular and adaptable to other geopolitical topics, the US-China trade focus simply serves as a high-impact use case to demonstrate its analytical value.

A. Data Preprocessing

To transform the raw articles into a usable format for event extraction and graph construction, we apply a structured preprocessing pipeline. This process is motivated by the need to reduce redundancy, support event-level granularity, and remove irrelevant noise that may interfere with downstream models.

1) Deduplication: We remove duplicate articles, including those that are syndicated across both news platforms. This step ensures that repeated reporting on the same events does not bias event clustering or inflate frequency-based metrics.
2) Segmentation: For sentence-level models such as AMR-guided and Text2Event, we segment each article into paragraphs and sentences. This enables localized event extraction, as most event mentions are contained within individual sentences or short spans.
3) Text Cleaning: We remove non-content elements such as HTML tags, email addresses, phone numbers, and other formatting artifacts. This step improves tokenization quality and prevents the models from assigning attention to irrelevant text.

The final dataset consists of 7,680 unique and cleaned articles. This corpus serves as the foundation for all subsequent modules in our pipeline, including event extraction using large language models, clustering of semantically similar event mentions, and inference of temporal and causal links between events.

III. Methodology

A. Event Extraction

In this work, we define an event as a structured unit that includes a trigger (usually a verb or noun phrase) and a set of arguments. These arguments may include agents, objects, locations, and temporal expressions [1]. Events may also include temporal or causal links to other events. This definition is consistent with widely used frameworks in event extraction and temporal annotation. The objective of event extraction is to convert unstructured news text into structured, semantic representations that support clustering and relation inference.

1) AMR-Guided Graph Encoding: As a baseline, we evaluate an event extraction model based on Abstract Meaning Representation (AMR) [2]. This method combines semantic parsing with contextual embeddings to identify event structures. Each sentence is parsed into an AMR graph using an existing semantic parser. A RoBERTa-based encoder is used to generate contextual token embeddings in parallel. Event triggers and entity mentions are aligned to their corresponding nodes in the AMR graph, forming a heterogeneous graph that captures both syntactic and semantic information. The model then refines this graph using attention-based message passing. A feedforward classifier is used to detect entities and triggers, and a hierarchical decoder, guided by the AMR graph, predicts argument spans and their roles.

Although this model uses a well-defined semantic framework, it has several limitations. It sometimes assigns incorrect entity types, such as labeling "market" as a geopolitical entity instead of an economic term. Trigger detection is often incomplete, especially for multi-word expressions. For instance, it may label only "war" as the trigger in the phrase "trade war". While some arguments are identified correctly, their roles are often too general or not well-matched to context. More importantly, the model does not capture relationships across sentences and cannot infer causal connections that are not explicitly stated, which limits its use in processing complex news narratives.

2) Text2Event: We also evaluated Text2Event [3], a transformer-based model that adapts the sequence-to-sequence generation paradigm to the event extraction task. It comprises two components: (1) a standard encoder-decoder transformer-based network and (2) a constrained decoding algorithm. Given a raw sentence, the decoder produces a linearized text representation of the structured event, preserving the underlying hierarchical information. To prevent the model from producing invalid formats of the output, e.g. invalid event types, mismatch arguments, or incomplete structure, a novel decode algorithm is deployed to replace the standard greedy approach.

We tested a pretrained version of Text2Event using models available through the OmniEvent [4] framework. The model is trained on the ACE 2005 dataset, which defines a limited set of eight event types, including Conflict, Transaction, and Justice. This restricted schema makes the model less suitable for open-domain settings such as geopolitical news, where events span a broader range of categories. Adapting the model to our context would require domain-specific fine-tuning, which involves significant annotation and computational effort.

In practice, the model shows limitations similar to the AMR-based approach. Trigger identification is sometimes incomplete, particularly for multi-word expressions. Argument roles, while often correctly extracted, tend to be generic and may not match the event context precisely. These issues limit the model's ability to extract high-quality event representations in complex real-world narratives without additional customization.

3) LLM-Based Event Extraction: To address the limitations of AMR-guided and sequence-to-structure models, we use a large language model (LLM) for event extraction. We frame the task as a structured question-answering problem. A generative language model, such as GPT-4 [5], is prompted to identify key events in each document, assign them to high-level event categories (such as Economic Warning or Trade Policy), and extract relevant arguments, including agents, objects, time expressions, and locations. The model is also prompted to detect potential relations between events, such as causal or follow-up links, and to provide short natural language summaries for each event.

This approach produces structured event representations that are both semantically rich and context-sensitive. It requires minimal supervision and captures not only direct mentions but also implied relationships. For example, in a document about tariff policies, the model successfully identified a warning by the Federal Reserve, a government action imposing tariffs, and a later diplomatic agreement that suspended them.

Compared to AMR-based methods, the LLM-based approach is more flexible. It handles complex sentence structures, correctly identifies multi-word event triggers, and works across domains without fine-tuning. While the model mainly captures intra-document relations, its structured outputs provide a strong basis for later stages such as clustering and relational inference.

B. Event Clustering

Since the dataset comprises articles from multiple news outlets and sources, it is common for the same real-world event to be mentioned multiple times, often with varied wording or emphasis. To avoid redundant representations in the knowledge graph and improve the quality of downstream inference, we implement an event deduplication pipeline that clusters event mentions referring to the same underlying event.

Each event is first transformed into a canonical textual representation. This representation is constructed by concatenating the event type, trigger phrase, and event summary into a single string. The goal of this canonicalization step is to standardize the format and provide consistent semantic input for encoding.

Next, we encode each canonicalized event into a dense embedding using a pretrained language model [6]. This model has been shown to produce high-quality semantic embeddings suitable for semantic similarity tasks.

Given the high dimensionality and scale of the data, conventional clustering algorithms such as K-Means are not computationally feasible. Instead, we employ Product Quantization (PQ) [7] for efficient approximate nearest neighbor search. PQ allows scalable similarity comparisons by quantizing the embedding space into sub-vector codebooks, significantly reducing memory and computation while preserving semantic relationships.

To determine whether two events should be assigned to the same cluster, we compute the cosine similarity between their embeddings. A similarity threshold is then applied to decide whether the pair should be linked. The threshold is a tunable hyperparameter that controls the granularity of clustering. Lower thresholds yield tighter clusters with higher precision, while higher thresholds increase recall but risk merging semantically distinct events.

Finally, we use a disjoint-set data structure [8] to perform transitive closure over the similarity links and form disjoint event clusters. This ensures that any chain of pairwise-similar events is merged into a single cluster, even if individual pairs are not directly connected. Each resulting cluster represents a single canonical event in the final event graph.

This clustering step is crucial for consolidating semantically equivalent event mentions across documents and sources. It enables the system to reason at the level of real-world events rather than isolated textual mentions, improving both the interpretability and structural integrity of the resulting knowledge graph.

C. Temporal and Causal Relations Extraction

The final stage of our pipeline focuses on identifying temporal and causal relationships between event clusters. While earlier steps extract discrete events and group co-referent mentions, this stage focuses on modeling how different events evolve and influence each other. Modeling these relationships is important for building coherent narratives that reflect real-world developments. For example, a policy announcement might trigger a change in trade talks, or new regulations could lead to supply chain disruptions. Identifying such links supports higher-level tasks like timeline construction, story tracking, and causal reasoning.

To guide this inference, we use intra-document event links as a weak supervisory signal. Each extracted event may include annotated relations to other events in the same document, with labels such as `RESPONSE_TO`, `INFLUENCED`, or `TRIGGERED`. These links are used to identify candidate relations between event clusters. Because inference is performed at the cluster level, where each cluster represents a single real-world event, we first map event-level relations to their corresponding clusters. Duplicate relations targeting the same cluster are removed to reduce noise.

To ensure consistent input for relation inference, we select a single representative event from each cluster. This representative is chosen based on the length of the event trigger or summary, giving priority to longer instances. Longer descriptions often carry more useful information, improving the accuracy of models like COMET and embedding-based similarity measures. By grounding relation inference on these representative events, we improve the precision and interpretability of the extracted links.

Commonsense-Based Relation Inference. Commonsense inference helps uncover implicit relationships that are not captured by annotated dependencies. To add plausible temporal and causal links to the event graph, we use a generative model trained on the ATOMIC knowledge base, following COMET-ATOMIC 2020 [9]. This model allows inference of likely causes, consequences, or follow-up events based on prior world knowledge, going beyond explicit text-based clues.

For each representative event, we create natural language prompts that match ATOMIC relation types, such as *xEffect* (likely consequence) and *isAfter* (typical follow-up event). These prompts are passed to the COMET model to generate possible continuations. The generated outputs are used as candidate relations between event clusters in the graph.

This approach addresses two main challenges. First, annotated dependencies are usually limited to events within the same document and may miss broader cross-document links. Commonsense generation helps extend inference to such cases. Second, it enables reasoning about events that are not well-covered in the training data or are expressed indirectly in the text. This is useful in domains where labeled causal or temporal links are scarce.

However, COMET may generate vague or unrelated outputs. To filter low-quality inferences, we apply a semantic validation step using Sentence-BERT [10]. For each candidate relation, we compute cosine similarity between the COMET-generated text and all representative event summaries. A relation is kept only if its similarity with a target summary exceeds a fixed threshold.

This filtering serves two purposes. It ensures that generated relations are consistent with the content of the dataset and removes unrelated or hallucinated outputs. As a result, the final graph contains high-confidence relations that are supported both by commonsense inference and semantic similarity. This method provides a scalable and weakly-supervised way to infer relations in settings without annotated labels.

IV. Knowledge Graph Construction & Analysis

A. Event Influence Analysis Using PageRank and Structural Clustering

Following the results of the events clustering and the causal relations extraction, we applied the PageRank algorithm to identify structurally central events, which frequently appear in the causal paths of multiple narratives. Notably, events such as "China banned all seafood imports" (PageRank: 0.0170) and "60% tariff on Chinese goods" (PageRank: 0.0122) received high scores, indicating their influence in anchoring broader narratives across domains such as trade, diplomacy, and security.

In parallel, we computed betweenness centrality to identify mediating events that act as narrative bridges by linking otherwise disconnected clusters. The event "60% tariff on Chinese goods" stood out, with both a high PageRank score (0.0122) and the highest betweenness centrality (0.0088). This dual prominence highlights its role as both a central influencer and a key connector within the event network.

A comparison of the top-ranked events based on both metrics is presented below (Table I).

TABLE I: Top Events by PageRank and Betweenness Centrality

ID	Event	PR	BC
17	**60% tariff on Chinese goods**	**0.0122**	**0.0089**
51	Military and political pressure	0.0041	0.0048
90	US-China tensions are high	0.0101	0.0031
178	War games around the island	0.0050	0.0026
150	Steep tariff increases	0.0064	0.0023
69	**China banned seafood imports**	**0.0170**	0.0020
3	Sweeping tariffs on goods	0.0068	0.0018
109	Military exercises ramped up	0.0041	0.0015
68	**South China Sea disputes**	**0.0126**	0.0014
148	Trade war threat	0.0091	0.0014

To further illustrate these insights, we present a zoomed-in view of the clustered event graph, focusing on the trade and tariff-related cluster where many of the high-ranking events are concentrated. As shown in Fig. 1, node size reflects PageRank scores,and colors represent clusters centered on key trigger events. Semantic labels were used to help interpret the thematic focus of each cluster.

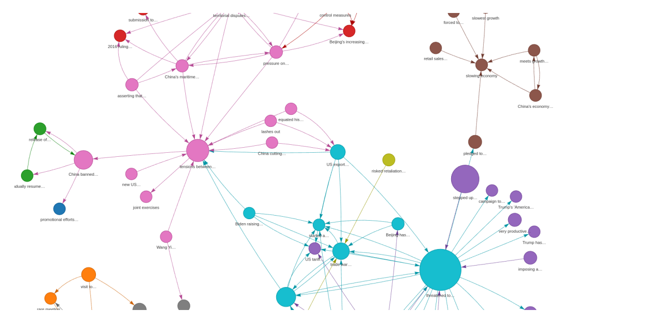


Fig. 1: Zoomed-in cluster visualization of trade-related influential events. Node size represents PageRank; colors indicate structurally defined clusters.

Structurally central and narratively bridging events such as the 60% tariff threat are positioned at the intersection of multiple clusters. These events play pivotal roles in transitioning narratives across economic and diplomatic domains. A full interactive version of the event graph is provided as a supplementary file, named `event_clusters_pyvis.html`, which enables deeper exploration of the event network and causal dynamics.

This integrated approach, which combines graph-based centrality metrics with structurally defined clustering, helps analysts and decision-makers anticipate ripple effects more effectively and navigate evolving geopolitical narratives with improved structural insight.

B. Interpretation and Business Value of the Event Knowledge Graph

Beyond constructing a causal-temporal event graph, the real value of this project lies in how the graph can be used to support strategic decision-making, risk monitoring, and policy analysis.

Our event graph enables structured exploration of how key geopolitical developments evolve over time and interact across domains such as trade, military, and diplomacy. To operationalize these insights, we present three complementary visualizations: the macro-level cluster graph, the causal flowchart, and the narrative tree. These are supported by centrality metrics (PageRank and betweenness) to highlight structurally important events.

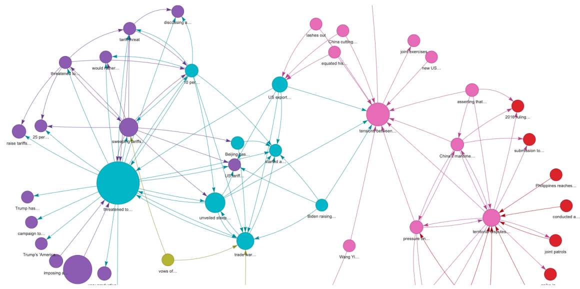


Fig. 2: Knowledge graph showing clustered event communities.

In Figure 2, we observe distinct event clusters representing key geopolitical and economic themes:

- Teal nodes capture narratives focused on trade and tariff tensions between the US and China.
- Pink nodes represent diplomatic interactions, including statements, retaliations, and formal meetings.
- Red nodes are associated with security flashpoints, including maritime disputes and joint military activities.
- Purple nodes capture political campaign rhetoric and policy signaling (e.g., "Trump has...", "imposing a...").

Several high-impact nodes act as narrative bridges or escalation triggers:

- The event "threatened to slap a 60 per cent flat fee on Chinese goods" sits at the intersection of purple and teal regions. It has both the highest PageRank and betweenness centrality, indicating it is central in many causal pathways and connects political rhetoric to concrete trade policy outcomes.
- "Sweeping tariffs" and "unveiled steep tariffs" are shown as large teal nodes with many outbound links, revealing their role as triggers for subsequent retaliation or policy escalation.
- The node "tensions between the US and China", located in the pink diplomatic cluster, connects to both trade and military events, reinforcing its role as a narrative aggregation point across domains.
- The cluster of red nodes surrounding "territorial disputes" and "joint patrols" captures regional tensions in the South China Sea. These security-related events are isolated from trade policy at times but are pulled into the graph via bridge nodes like "asserting maritime rights".

These colored clusters and their interconnections provide an intuitive map of geopolitical entanglement. Analysts can trace how:

- Economic policies spark security responses (e.g., from teal to red clusters).
- Campaign rhetoric becomes actionable policy, moving from purple to teal.
- Diplomatic actions attempt to defuse or reframe escalating tensions by linking pink nodes to all others.

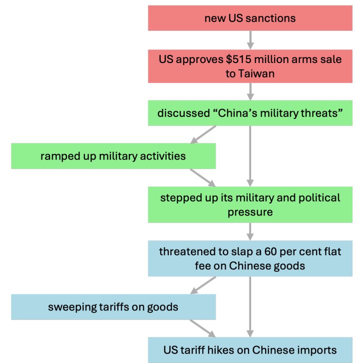


Fig. 3: Causal flowchart outlining a trade-policy escalation sequence, starting from US sanctions and progressing into economic and military reactions.

Figure 3 complements the structural knowledge graph by isolating a narrative slice of causally connected events. The flow begins with a diplomatic trigger — "new US sanctions" — which catalyzes a series of developments spanning multiple domains. The approval of a major arms sale to Taiwan prompts a military response from China ("ramped up military activities"), which then escalates into broader pressure campaigns and retaliatory economic measures ("60 per cent tariff threat", "sweeping tariffs on goods").

Each colored node in the figure corresponds to a distinct policy domain:

- Red nodes denote diplomatic or policy declarations.
- Green nodes represent military movements and rhetoric.
- Blue nodes reflect trade and economic responses.

This flowchart illustrates how cross-domain escalation unfolds as a chain reaction, offering clear value for multiple strategic applications. In scenario planning, analysts can anticipate downstream impacts triggered by early-stage diplomatic moves, enabling better preparedness for subsequent developments. For crisis mapping, stakeholders can trace the ripple effects of a single policy decision through interconnected economic and military consequences, identifying critical escalation points. The structure also supports strategic communication by helping policymakers convey the stakes and trajectories of escalation in a clear and intuitive linear format, making complex geopolitical developments more accessible to both decision-makers and the public.

Figure 4 extends our analysis by presenting a branching narrative tree that captures multiple causal trajectories from key trigger events. In this view, events like "large-scale war games around Taiwan" serve as roots from which several distinct escalation paths emerge. These include sequences related to military buildup (e.g., "staged two rounds," "held two-day drills"), diplomatic signaling (e.g., "China had held..."), and economic retaliation (e.g., "Beijing vowed retaliation,"

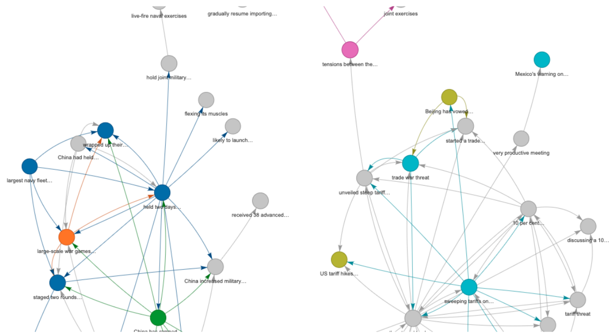


Fig. 4: Narrative tree showing escalation paths from "large-scale war games" and "tariff threats."

"unveiled steep tariffs"). This branching format complements the structural clusters and linear flowchart by revealing the diversity of responses that can arise from a single originating event.

While the cluster graph emphasized network influence and the flowchart traced escalation timelines, the narrative tree provides a scenario-based perspective. It allows analysts to simulate divergent paths, identify which events lead to broader ripple effects, and assess the likelihood of cross-domain propagation. For policymakers and strategic planners, this format supports contingency planning by highlighting key points of potential divergence. These are moments where a single trigger can escalate into entirely different geopolitical outcomes, depending on the responses of involved parties. Combined with the prior visualizations, the tree further reinforces the analytical value of the knowledge graph in anticipating, explaining, and navigating complex event dynamics.

V. Experiments and Results

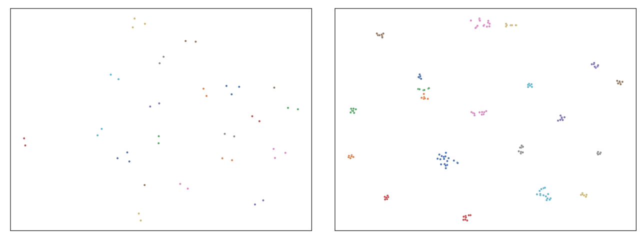
A. Event Clustering Evaluation

Cluster quality is measured using the silhouette coefficient [11], which reflects the balance between intra-cluster cohesion and inter-cluster separation. Table II reports silhouette scores for both models across a range of similarity thresholds. BGE M3 consistently achieves higher silhouette scores than BGE v1.5 at corresponding thresholds, suggesting that it produces more compact and well-separated clusters. The highest silhouette score (0.9289) is obtained by BGE M3 at a threshold of 0.05. In general, lower thresholds result in higher silhouette scores, likely due to the stricter criteria for grouping, which reduces the likelihood of merging unrelated events.

Threshold	Bge v1.5	Bge M3
0.05	0.7922	0.9289
0.1	0.5510	0.5926
0.2	0.4135	0.4669
0.5	0.4502	0.3838
0.75	0.0000	0.2639

TABLE II: Silhouette coefficients using two embedding models with various similarity threshold

To better understand the effects of similarity thresholding, we visualize the top 20 most frequently mentioned events under thresholds of 0.05 and 0.1 (Figure 5). At 0.05, clusters are tightly grouped and contain only a few event mentions, typically between 1 and 3. This reflects strong cohesion and low noise, but the conservative thresholding can lead to under-clustering, where semantically equivalent events remain unmerged due to minor surface differences. At threshold 0.1, clusters become more dispersed but include a greater number of mentions, indicating improved recall and better coverage of paraphrased or variably phrased duplicates. However, this comes at the cost of slightly reduced cohesion and a modest drop in silhouette score.



(a) Threshold = 0.05 (b) Threhsold = 0.1

Fig. 5: t-SNE visualization of top 20 most mentioned events using different threshold.

These results highlight several limitations of the silhouette score when used as the primary clustering metric in this context. While the score is useful for evaluating geometric separation, it does not directly assess whether true duplicate or coreferent events are clustered together. For example, two mentions of the same event expressed in different styles may be placed in separate clusters, yet still yield a high silhouette score due to clear embedding separation. This discrepancy suggests that silhouette score may overestimate clustering quality in cases where lexical diversity masks semantic equivalence.

Additionally, poor clustering can impact downstream components in the pipeline. If semantically identical events are not merged, the resulting graph becomes fragmented, with multiple disconnected nodes representing the same real-world occurrence. This can reduce interpretability and affect the quality of relation inference.

To address these limitations, future work may incorporate auxiliary signals during clustering, such as argument-level similarity or named entity overlap. Weak supervision or contrastive fine-tuning of the embedding space may also help bring semantically equivalent but lexically diverse mentions closer together. Incorporating these strategies could lead to more robust and semantically grounded event clustering that aligns better with coreference and real-world coherence.

B. Temporal and Causal Relations Results

Evaluating temporal and causal relations between events is challenging due to the lack of labeled ground truth. Supervised benchmarks for this task are limited, especially in open-domain or geopolitical news domains. As a result, standard metrics such as precision, recall, and F1-score are not applicable in our setting. To address this, we use a human evaluation protocol to assess the quality and interpretability of the inferred relations.

We conduct a manual evaluation on 60 sampled relations between event clusters across four dimensions:

- Plausibility: Is the relation contextually and logically reasonable based on commonsense or real-world knowledge?
- Directionality: Does the inferred direction correctly reflect causal or temporal precedence?
- Explanation Coherence: Is the COMET-generated explanation semantically consistent with the relation?
- Relation Type Accuracy: Does the predicted label (e.g., causal or temporal) correctly describe the link?

These criteria allow us to evaluate both the semantic correctness of individual links and the overall consistency of the event graph. The evaluation results are shown in Table III.

Metric	Score
Relation Plausibility	45/60 (75.00%)
Direction Correctness	45/60 (75.00%)
Explanation Coherence	43/60 (71.67%)
Relation Type Accuracy	44/60 (73.33%)

TABLE III: Manual evaluation results of inferred temporal and causal relations.

The evaluation results show that the proposed method generates event relations that are generally meaningful and contextually appropriate. About 75% of the inferred relations were rated as plausible, and a similar proportion were judged to have the correct direction. These results suggest that combining COMET-based reasoning with SBERT-based filtering produces coherent temporal and causal links across event clusters.

The explanation coherence score was slightly lower, at 71.7%, indicating that while many generated explanations align with the inferred relations, some lack specificity or relevance. Relation type accuracy was 73%, though some confusion occurred between causal links and weaker associative connections.

These results highlight several strengths of the approach. Most inferred relations align with commonsense knowledge, capturing cause-effect patterns and temporal order in a way that reflects real-world logic. COMET's generative outputs also help identify implicit links that are not directly stated in the text, which is useful in domains like geopolitics or finance where causality is often indirect. The method performs well across both causal and temporal categories, showing generalization without the need for domain-specific rules or labeled supervision.

Despite the encouraging results, the evaluation highlights several limitations. A key issue lies in the coherence of COMET-generated explanations. In some instances, the model produces vague, speculative, or semantically misaligned outputs that do not correspond well with the target event. These inconsistencies can introduce ambiguity into the inferred relations and weaken the reliability of the resulting event graph.

Directionality errors are also common, especially when event descriptions include multiple entities or ambiguous references. For example, the model may assign the wrong causal or temporal direction when handling pronouns or overlapping actions. In addition, COMET sometimes produces hallucinated content, where the events or motivations are not present in the source or target summaries. These errors can distort the predicted relation type, such as mistaking a temporal sequence for a causal link, which affects the interpretability and reliability of the graph.

Several improvements can help address these issues. First, adding entity-aware event representations through techniques such as coreference resolution or named entity linking may reduce ambiguity and improve directionality accuracy. Second, a small classifier could be used to remove COMET outputs that are semantically inconsistent or out-of-domain. Third, incorporating more linguistic cues such as temporal expressions, modal verbs, or discourse markers may improve the model's ability to distinguish between causal and temporal relations.

VI. Discussion

This work presents a structured approach for extracting event representations from unstructured news text, with a focus on capturing interactions between distinct events. The framework emphasizes three main components: detecting event triggers, clustering event mentions, and inferring temporal and causal links across clusters. These steps allow the system to model narrative structure with limited supervision and without relying on fixed ontologies or annotated training sets.

The current design, however, abstracts away from other important aspects of event understanding. It does not model entities in detail, distinguish between fine-grained event types, or construct global timelines. While these choices simplify the framework, they also limit its utility for applications such as knowledge graph construction and multi-hop reasoning. These gaps define the scope of the current work and suggest clear directions for future research.

A. *Absence of Structured Entity Representations*

While entities such as people, organizations, and locations are central to event semantics, the current framework does not model them explicitly. Entities are extracted as arguments of events but are not represented as standalone nodes, nor are they resolved or linked across documents. As a result, mentions like "United States," "US," and "America" are treated as different entities, even though they refer to the same actor. Similarly, phrases such as "Federal Reserve officials" and "the central bank" may not be aligned, which limits the system's ability to track continuity across events.

This absence of coreference resolution, entity disambiguation, and linking creates challenges for building a structured and queryable event graph. Without identifying and grounding entity mentions to external sources, it is difficult to support tasks like temporal reasoning, actor-level tracking, or analysis across multiple documents. Addressing this issue would require integrating entity linking tools based on resources like Wikidata or DBpedia, along with cross-document coreference resolution methods. These additions would enable canonicalization of entity mentions and support more coherent graph construction. They would also expand the system's scope from event-event inference to entity-centered analysis, which is essential for scalable and interpretable knowledge representation.

B. Lack of Ontology Alignment and Event Typing

The current framework assigns event type labels such as Economic Warning, Trade Policy, and Diplomatic Agreement as part of its structured output. However, these labels are generated heuristically or via LLM prompting and are not aligned with any standardized event ontology. As a result, the event types lack semantic consistency and interoperability with established schemas like FrameNet, ACE, or other domain-specific ontologies.

This limitation affects how well the system can organize and compare events across documents or domains. Without a standardized event schema, it is challenging to aggregate or compare events across sources, domains, or systems, particularly in cases that require taxonomic organization or cross-document consistency. Additionally, the absence of structured semantic roles reduces the graph's utility for downstream applications that require a detailed understanding of event subtypes and argument roles. Future work could incorporate supervised classification methods using established event ontologies or apply unsupervised schema induction to uncover latent event structures. These improvements would improve the expressiveness and compatibility of the extracted graph, making the graph more suitable for integration into larger knowledge systems.

C. Limitations in Global Temporal Reasoning

The current framework infers temporal and causal relations between event clusters through pairwise comparisons, identifying directional links such as causes and temporal precedence. These links can connect events both within a single document and across multiple documents, and are represented as edges in the event graph. However, each relation is inferred independently, with no mechanism for modeling transitive dependencies or ensuring global temporal consistency across event sequences. As a result, while the graph captures local temporal structures, it does not generate unified, global timelines or support reasoning over multi-step event chains.

Future work should focus on extending the framework to model global temporal dependencies. One approach could involve graph-based inference to handle long-range temporal relationships and maintain consistency across event chains. Another option is to use timeline induction techniques to generate ordered chronologies that go beyond pairwise links. These enhancements would improve the system's ability to track narratives across documents and support downstream tasks like cross-document summarization and temporal reasoning in knowledge graphs.

VII. Conclusion

We present a modular framework for structuring unstructured news articles by extracting events, clustering related mentions, and identifying temporal and causal links between them. The framework combines large language models for zero-shot event extraction, unsupervised methods for event clustering, and weakly supervised approaches for relation inference. This approach produces an event-centric graph that captures narrative dynamics across documents and supports structured analysis of event interactions. Unlike many existing methods, the framework does not rely on predefined schemas or domain-specific fine-tuning, making it adaptable to open-domain and evolving contexts such as geopolitical reporting.

The system enables a range of downstream applications, including event tracking, policy impact analysis, and knowledge graph construction. Beyond basic structuring, we incorporate graph-based influence analysis to identify central and bridging events using PageRank and betweenness centrality. This helps reveal which events anchor or connect major narrative themes, providing insight into how stories evolve across domains such as trade, diplomacy, and security. An interactive visualization of the event graph further supports exploratory analysis.

Although we do not quantitatively evaluate event extraction in isolation, its quality is demonstrated through qualitative inspection and effective integration with later components. The resulting graph provides an interpretable foundation for event-level reasoning and supports higher-level analytical workflows.

Future work should focus on extending the framework with entity-level modeling, alignment with existing event ontologies, and improved global temporal reasoning. Additions such as multi-hop inference, global timeline construction, and cross-document entity linking will further improve the system's applicability in tasks that require long-range consistency and multi-source analysis.

As of April 2025, the global trade landscape has been significantly impacted by the Trump administration's implementation of sweeping tariffs. These measures include a baseline 10% tariff on all imports and a substantial 145% tariff specifically targeting Chinese goods. In retaliation, China has imposed a 125% tariff on U.S. imports, leading to a marked decline in bilateral trade and escalating economic tensions between the two nations. These developments have introduced considerable volatility into global markets and have raised concerns about the stability of international trade relations.

While our framework was initially developed and tested using data from February 2024 to February 2025, its modular and scalable design ensures adaptability to evolving geopolitical events. The system's reliance on zero-shot event extraction, unsupervised clustering, and weakly supervised inference allows for the seamless integration of new data, enabling real-time analysis of current developments such as the ongoing U.S.-China trade tensions. This adaptability underscores the framework's utility in providing structured, interpretable insights into complex, dynamic geopolitical narratives.

References

[1] Linguistic Data Consortium. ACE (Automatic Content Extraction) 2005 Multilingual Training Corpus, 2005. https://catalog.ldc.upenn.edu/LDC2006T06.

[2] Zixuan Zhang and Heng Ji. Abstract meaning representation guided graph encoding and decoding for joint information extraction. In *Proc. The 2021 Conference of the North American Chapter of the Association for Computational Linguistics-Human Language Technologies (NAACL-HLT2021)*, 2021.

[3] Yaojie Lu, Hongyu Lin, Jin Xu, Xianpei Han, Jialong Tang, Annan Li, Le Sun, Meng Liao, and Shaoyi Chen. Text2event: Controllable sequence-to-structure generation for end-to-end event extraction. *arXiv preprint arXiv:2106.09232*, 2021.

[4] Xin Jin, Yujie Zhang, Wanjun Wang, Junyi Jessy Li, and Heng Ji. OmniEvent: A comprehensive event extraction benchmark for multi-lingual and cross-lingual scenarios. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7455–7471, 2022.

[5] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[6] Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation, 2023.

[7] Herve Jegou, Matthijs Douze, and Cordelia Schmid. Product quantization for nearest neighbor search. *IEEE transactions on pattern analysis and machine intelligence*, 33(1):117–128, 2010.

[8] John E. Hopcroft and Jeffrey D. Ullman. Set merging algorithms. *SIAM Journal on Computing*, 2(4):294–303, 1973.

[9] Jena D Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. (comet-) atomic 2020: On symbolic and neural commonsense knowledge graphs. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 6384–6392, 2021.

[10] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

[11] Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

Appendix

Appendix A: Breakdown of Workload

Task		**Person-In-Charge**
Data Collection		Celine
Data Preprocessing		Celine
Event Extraction	Text2Event (OmniEvent)	Khoa
	AMR-Guided Graph Encoding	Clare
	LLM-Based	Celine, Thet Su
Event Clustering		Khoa
Temporal & Causal Relation Extraction		Clare
Knowledge Graph & Inference		Celine, Thet Su
Final Report		All